

JGI Microbial Single Cell Program

Single Cell Data Decontamination

Despite our best efforts, it is possible that there are some contigs in your single cell genome that are from contaminant organisms. In addition to external contaminants, there may be contaminants from within your particular sample in the form of free DNA that made it into the well along with your single cell. All single cell genomes sequenced at the JGI go through an automated contaminant removal pipeline. In order to prevent contaminant data from propagating in the databases to cause errors in the annotation of future genomes/metagenomes, the automated system is tuned to be aggressive in its contaminant removal. This means that it is likely that some contigs that were truly from your genome have been removed. Because these contigs that were removed may be of interest to you, we provide two versions of each single cell genome: a screened version that has gone through our automated pipeline and is added to the databases for use in annotation of future genomes/metagenomes, and an unscreened version that contains all the contigs, but is not part of the databases used for annotation. If you do not want to rely on the aggressively cleaned version that we provide, you may do a manual contamination removal using the unscreened dataset. While there are no clear rules on the identification and removal of contamination (i.e. phage or horizontal gene transfer may be difficult to discriminate from contamination), we would like to provide some recommendations and guidance.

Because your target genome plus the contaminant sequences are essentially a small metagenome, and the tools useful for analyzing metagenomes are the ones we use for the contamination screening, you will initially log into the IMG/MER system and not IMG/ER.

After logging into the IMG/MER system you need to find your genomes.

Under the Find Genomes tab you can search for any genome in IMG by a variety of criteria.

JGI Joint Genome Institute

ALL Genomes Hi Scott Clingenpeel | Logout (JGI SSO)

img/mer

INTEGRATED MICROBIAL GENOMES with MICROBIOME SAMPLES

EXPERT REVIEW

Find Genomes | Find Genes | Find Functions | Compare Genomes | Analysis Cart | OMICS | ABC | My IMG | Data Mats | Using

IMG/MER Content

Dataset	Count
Bacteria	24039
Archaea	501
Eukarya	190
Plasmids	1186
Viruses	3882
Genome Fragments	1188
Metagenome	2407
Total Datasets	35018
My Private Datasets	10
GEBAs	256

Last Genome updated: 2015-02-22
Last Sample updated: 2015-02-24

[Metagenome Projects Map](#)
[System Requirements](#)

[Hands on training available at the Microbial Genomics & Metagenomics Workshop](#)

The Integrated Microbial Genomes (IMG) system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The IMG data warehouse integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available isolate and single cell genomes and a rich set of publicly available metagenome samples.

IMG/MER (Nucleic Acids Research Volume 42 Issue D1) provides users with tools (IMG/MER Map) for analyzing their private (password protected access) metagenome samples in the context of all public (free access) genome and metagenome samples in IMG.

[IMG/MER Statistics](#) [Data Submission Site](#)

IMG/MER contains 245 public studies, 3407 public metagenome datasets (3194 unique samples) distributed as follows:

Engineered	205	Environmental	2048	Host-associated	1154
Bioreactor	16	Air	31	Annelida	34
Bioremediation	21	Aquatic	1209	Arthropoda	75
Biotransformation	26	Terrestrial	808	Birds	5
Food production	3			Cnidaria	2
Lab enrichment	18			Human	861
Solid waste	23			Mammals	27
Wastewater	98			Microbial	3
				Mollusca	9
				Plants	122
				Porifera	8
				Tunicates	8

2015-03-17-08:00:02

Contact Us
[Accessibility/Section 508](#)
[Disclaimers](#)
Version 4.510 Oct 2014
genweb09 genome1_shared 5.010000 2015-03-16-13:39:52 128.3.91.138
© 1997-2015 The Regents of the University of California

Office of Science

Home > Find Genomes

2 genomes retrieved.

Genome Field Search Results

hint: Go to [Preferences](#) to show or hide plasmids, GFragment and viruses.
Go to home page statistics under [IMG Genomes](#) to select individual phylogenetic domains or all genomes.

Domains(D): *=Microbiome,
B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, G=GFragment, V=Viruses.
Genome Completion(C): F=Finished, P=Permanent Draft, D=Draft.

Add Selected to Genome Cart Select All Clear All

Filter column: Domain Filter text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Domain	Status	Genome Name / Sample Name	Study Name	Sequencing Center	Genome Size	Gene Count
<input type="checkbox"/>	B	P	Colwellia sp. SCGC AC281-C05 (contamination screened)	Single cell sequencing of Cycloclasticus and Colwellia ecotypes from the Deepwater Horizon oil spill	DOE Joint Genome Institute (JGI)	888216	822
<input type="checkbox"/>	B	P	Colwellia sp. SCGC AC281-C05 unscreened y2	Single cell sequencing of Cycloclasticus and Colwellia ecotypes from the Deepwater Horizon oil spill	DOE Joint Genome Institute (JGI)	1057901	1023

Export Page 1 of 1 << first < prev 1 next > last >> All

Two versions of each single cell genome.

Each of your single cell genomes should have two versions in IMG: A “contamination screened” version that was decontaminated with JGI’s automated pipeline, and an “unscreened” version that contains all the contigs greater than 2kb in length. Click on the name of the unscreened genome in order to bring up the genome overview page. Scroll down to the Genome Statistics.


Genome Statistics

hint: To view rows that are zero, go to [MyIMG preferences](#) and set “Hide Zeroes in Genome Statistics” to “No”.

	Number	% of Total
DNA, total number of bases	1057901	100.00%
DNA coding number of bases	930004	87.91%
DNA G+C number of bases	384395	36.34% ¹
DNA scaffolds	82	100.00%
Genes total number	1023	100.00%
Protein coding genes	1003	98.04%
RNA genes	20	1.96%
rRNA genes	4	0.39%
5S rRNA	2	0.20%
16S rRNA	1	0.10%
23S rRNA	1	0.10%
RNA genes	13	1.27%
Other RNA genes	3	0.29%
Protein coding genes with function prediction	798	78.01%
without function prediction	205	20.04%
Protein coding genes with enzymes	299	29.23%
w/o enzymes but with candidate KO based enzymes	22	2.15%
Protein coding genes connected to Transporter Classification	103	10.07%
Protein coding genes connected to KEGG pathways ³	300	29.33%
not connected to KEGG pathways	703	68.72%
Protein coding genes connected to KEGG Orthology (KO)	571	55.82%
not connected to KEGG Orthology (KO)	432	42.23%
Protein coding genes connected to MetaCyc pathways	254	24.83%
not connected to MetaCyc pathways	749	73.22%
Protein coding genes with COGs ³	693	64.81%
with KOGs ³	186	18.18%
with Pfam ³	828	80.94%
with TIGRfam ³	402	39.30%
with InterPro	540	52.79%
with IMG Terms	203	19.84%

Click here to bring up a list of all your scaffolds.

Select all the scaffolds
and then add them to
your Scaffold Cart.

img/mer  **INTEGRATED MICROBIAL GENOMES**
EXPERT REVIEW with MICROBIOME SAMPLES


IMG/ER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Marts Using

[Home](#) > Find Genomes Loaded.

Chromosome Viewer

Scaffolds and contigs for [Colwellia sp. SCGC AC281-C05 unscreened v2](#)


User Selectable Coordinates

Filter column: Scaffold Filter text: 

Page 1 of 1 << first < prev 1 next > last >> All

Column Selector

Select	Scaffold	Length (bp)	GC	Type	Topology	Read Depth	No. Genes	Coordinate Range
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_10_len_30900_cov_1435_18_ID_2761.48	30900	0.36	genomic DNA	circular	1.00	24	1_30900
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_11_len_29121_cov_21_2434_ID_15.56	29121	0.37	genomic DNA	circular	1.00	24	1_29121
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_12_len_27820_cov_1860_65_ID_39.23	27820	0.36	genomic DNA	circular	1.00	26	1_27820
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_13_len_25426_cov_2308_3_ID_3933.60	25426	0.37	genomic DNA	circular	1.00	23	1_25426
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_14_len_25022_cov_11_551_ID_19.33	25022	0.36	genomic DNA	circular	1.00	19	1_25022
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_15_len_24180_cov_114_91_ID_3519.46	24180	0.37	genomic DNA	circular	1.00	24	1_24180
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_16_len_23849_cov_25_4328_ID_1113.34	23856	0.37	genomic DNA	circular	1.00	18	1_23856
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_17_len_21851_cov_65_6098_ID_95.19	21850	0.35	genomic DNA	circular	1.00	20	1_21850
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_18_len_20046_cov_85_9461_ID_25.29	20046	0.34	genomic DNA	circular	1.00	19	1_20046
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_19_len_19099_cov_29_7508_ID_67.13	19100	0.34	genomic DNA	circular	1.00	20	1_19100
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_1_len_87465_cov_100_697_ID_1.9	87448	0.36	genomic DNA	circular	1.00	77	1_87448
<input type="checkbox"/>	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_20_len_18882_cov_20_5883_ID_101.37	18883	0.36	genomic DNA	circular	1.00	18	1_18883

img/mer  **EXPERT REVIEW** **INTEGRATED MICROBIAL GENOMES** *with MICROBIOME SAMPLES*

IMG/MER Home Find Genomes Find Functions Compare Genomes **Analysis Cart** OMICS ABC My IMG Data Marts Using

Home > Analysis Cart 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis Phylogenetic Distribution

Add Genomes of Selected Scaffolds to Cart Add Genes of Selected Scaffolds To Cart

Toggle Selected Select All Clear All Remove Selected

Filter column: Scaffold ID Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_60_len_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	4	2079	0.30	
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_22_len_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_43_len_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	8	5026	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_37_len_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_40_len_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	9	5946	0.36	1

You can get back to the Scaffold Cart anytime by clicking Scaffolds under the Analysis Cart tab.

Click here to return to the genome overview page.


You will screen all of the scaffolds in your genome and put them in a clean data bin and a contaminants bin. A good place to start with contamination screening is to look at any ribosomal RNA sequences in your genome.

Genome Statistics

hint: To view rows that are zero, go to [MyIMG preferences](#) and set "Hide Zeros in Genome Statistics" to "No".

	Number	% of Total
DNA, total number of bases	1057901	100.00%
DNA coding number of bases	930004	87.91%
DNA G+C number of bases	384395	36.34% ¹
DNA scaffolds	62	100.00%
Genes total number	1023	100.00%
Protein coding genes	1003	98.04%
RNA genes	20	1.96%
rRNA genes	4	0.39%
5S rRNA	2	0.20%
16S rRNA	1	0.10%
23S rRNA	1	0.10%
tRNA genes	13	1.27%
Other RNA genes	3	0.29%
Protein coding genes with function prediction	738	78.01%
without function prediction	205	20.04%
Protein coding genes with enzymes	299	29.23%
w/o enzymes but with candidate KO based enzymes	22	2.15%
Protein coding genes connected to Transporter Classification	103	10.07%
Protein coding genes connected to KEGG pathways ³	300	29.33%
not connected to KEGG pathways	703	68.72%
Protein coding genes connected to KEGG Orthology (KO)	571	55.82%
not connected to KEGG Orthology (KO)	432	42.23%
Protein coding genes connected to MetaCyc pathways	254	24.83%
not connected to MetaCyc pathways	749	73.22%
Protein coding genes with COGs ³	663	64.81%
with KOGs ³	186	18.18%
with Pfam ³	828	80.94%
with TIGRfam ³	402	39.30%
with InterPro	540	52.79%
with IMG Terms	202	19.84%

Click here to bring up a list of all your rRNA genes.



INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

[IMG/MER Home](#)
[Find Genomes](#)
[Find Genes](#)
[Find Functions](#)
[Compare Genomes](#)
[Analysis Cart](#)
[OMICS](#)
[ABC](#)
[My IMG](#)

[Home](#) > [Find Genomes](#)
4 rRNA's, 0 tRNA's, 0 ncRNA's retrieved.

RNA Genes

Filter column: Gene ID Filter text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Gene ID	Locus Type	Gene Product Name	Gene Symbol	Coordinates	Length (bp)	Scaffold ID	Contig Length	Contig GC	Contig Read Depth
<input type="checkbox"/>	2599859935	rRNA	5S rRNA, Bacterial TSU	5S	45..159(-)	115	Ga0063151_NODE_38_len_6038_cov_324_761_ID_3919.12	6038	0.46	1.00
<input type="checkbox"/>	2599859936	rRNA	23S rRNA, Bacterial LSU	23S	376..3270(-)	2895	Ga0063151_NODE_38_len_6038_cov_324_761_ID_3919.12	6038	0.46	1.00
<input type="checkbox"/>	2599859938	rRNA	16S rRNA, Bacterial SSU	16S	3883..5440(-)	1558	Ga0063151_NODE_38_len_6038_cov_324_761_ID_3919.12	6038	0.46	1.00
<input type="checkbox"/>	2599860267	rRNA	5S rRNA, Bacterial TSU	5S	109..223(+)	115	Ga0063151_NODE_9_len_31955_cov_57_545_ID_13.38	31955	0.37	1.00


Export Page 1 of 1 << first < prev 1 next > last >> All

Add Selected to Gene Cart Select All Clear All

If the length of the gene is too short then it will not be phylogenetically informative.

Clicking the links in this column will allow you to retrieve the sequence of the gene.

This column tells you which scaffold each gene is located on.



INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

[IMG/MER Home](#)
[Find Genomes](#)
[Find Genes](#)
[Find Functions](#)
[Compare Genomes](#)
[Analysis Cart](#)
[OMICS](#)
[ABC](#)
[My IMG](#)

[Home](#) > [Find Genes](#)
Loaded

Gene Detail


[RNA Information](#)
[RNA Neighborhood](#)
[External Sequence Search](#)
[RNA Homologs](#)

RNA Information

Gene ID	2599859938
Gene Symbol	16S
Locus Tag	Ga0063151_00169
IMG Product Name	16S rRNA, Bacterial SSU
Original Gene Product Name	16S rRNA, Bacterial SSU
IMG Product Source	
Description	
Genome	Cobwellia sp. SCGC AC281-C05_v1558seq.v2
DNA Coordinates	3883..5440 (+)1558bp
Scaffold Source	Cobwellia sp. SCGC AC281-C05_Ga0063151_NODE_38_len_6038_cov_324_761_ID_3919.12 (6038bp)
IMG ORF Type	
GC Content	0.52
External Links	
Features	Name = "16S rRNA, Bacterial SSU" Note = "frame="

Add To Gene Cart

RNA Neighborhood



red = Current Gene
||||| CRISPR array
[Sequence Viewer For Alternate ORF Search](#)
Chromosome Viewer colored by [Select Function](#)

External Sequence Search

[NCBI BLAST](#)
[Green Genes BLAST](#)

Clicking here will bring up the sequence of the gene.

This indicates which scaffold the gene is on. Clicking it will take you to that scaffold.

This box shows the scaffold with all the genes on it. The gene you are currently examining is in red.

Clicking here allows you to BLAST the sequence in GenBank or 16S sequences can be compared to the GreenGenes database.

BLAST the rRNA sequences to see if they come from your target genome or from a contaminant.

Return to the Scaffold Cart under the Analysis Cart tab.

Select the scaffolds that contain rRNA genes from your target genome then go to the Upload Export & Save tab.

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE60_len_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	3	2079	0.30	1
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE22_len_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE43_len_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	6	5026	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE37_len_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE40_len_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	9	5946	0.36	1

Upload Scaffold Cart

You may upload a scaffold cart from a tab-delimited file. The file should have the column headers 'Scaffold ID'. (This file can be created by selecting [Scaffold Cart in Excel](#) button below.)

File to upload:
 No file chosen

Export Scaffold Data

You may export data for scaffolds selected in the cart.

Save Scaffolds to My Workspace

hint: Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save selected scaffolds to [My Workspace](#). (Special characters in file name will be removed and spaces converted to _)

☒ Save to File name:

☐ Append to the following scaffold set:

☐ Replacing the following scaffold set:

Chose a name for your bin of clean scaffolds and Save Selected to Workspace.

Go back to the Scaffold Cart and reselect the scaffolds that contain rRNA genes from your target genome then Remove Selected from the Scaffold Cart.

62 scaffold(s) in cart

Scaffolds in Cart | Upload & Export & Save | Function Profile | Histogram | Kmer Analysis | Phylogenetic Distribution

Add Genomes of Selected Scaffolds to Cart | Add Genes of Selected Scaffolds To Cart

Toggle Selected | Select All | Clear All | Remove Selected

Filter column: Scaffold ID | Filter: text | Apply

Export | Page 1 of 1 | << first < prev | next > last >> | All

Column Selector | Select Page | Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_60_len_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	3	2079	0.30	1
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_22_len_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_43_len_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	6	5026	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_37_len_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_40_len_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	9	5946	0.36	1

Follow the same procedure as above to select any scaffolds with contaminant rRNA genes, save them to a separate workspace and then remove them from the Scaffold Cart. At this time you may also want to save the remaining scaffolds in an “unchecked” workspace. When you leave IMG/MER the contents of your carts are not saved, but your workspaces are. You can get to the saved workspaces by going to Workspace under the My IMG tab.

My Workspace - Scaffold Sets

Scaffold Sets | Import & Export | Genomes & Genes | Function Profile | Histogram | Kmer Analysis | Phylogenetic Distribution

Select	File Name (click the link to each individual set)	Number of Scaffolds
<input type="checkbox"/>	clean	2
<input type="checkbox"/>	contaminants	4
<input type="checkbox"/>	unchecked	56

Add Selected to Scaffold Cart | Select All | Clear All | Remove Selected

My IMG

- MyIMG Home
- Annotations
- My Job
- Preferences
- Workspace
 - Gene Sets
 - Function Sets
 - Genome Sets
 - Scaffold Sets
 - Export Workspace
- Logout

Go back to your Scaffold Cart that now only contains unchecked scaffolds. Since the rest of the scaffolds do not contain any rRNA genes to serve as phylogenetic markers, their identity must be determined by looking at all the genes present on each scaffold.

img/mer **INTEGRATED MICROBIAL GENOMES**
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/ER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Marts Using

Home > Analysis Cart 60 scaffolds in cart

Scaffold Cart

60 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis **Phylogenetic Distribution**

Add Genomes of Selected Scaffolds to Cart Add Genes of Selected Scaffolds To Cart

Toggle Selected Select All Clear All Remove Selected

Filter column: Sequence Length (bp) Filter text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193596	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_1_len_87465_cov_100_697_ID_1.9	Colwellia sp. SCGC AC281-C05 unscreened v2	77	87466	0.36	1
<input type="checkbox"/>	2599193639	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_2_len_74585_cov_22_837_ID_3.52	Colwellia sp. SCGC AC281-C05 unscreened v2	66	74585	0.37	1
<input type="checkbox"/>	2599193607	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_3_len_68972_cov_143_421_ID_5.20	Colwellia sp. SCGC AC281-C05 unscreened v2	72	68972	0.36	1
<input type="checkbox"/>	2599193642	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_4_len_62163_cov_1481_54_ID_9.55	Colwellia sp. SCGC AC281-C05 unscreened v2	57	62163	0.37	1

To check if a scaffold is a contaminant, select the scaffold then click on Phylogenetic Distribution of Genes button.

Click on the Distribution by BLAST percent identities button.

img/mer **INTEGRATED MICROBIAL GENOMES**
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/ER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Marts Using

Home > Analysis Cart 60 scaffolds in cart

Scaffold Cart

60 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis **Phylogenetic Distribution**

Limit: Limit the numbers of scaffolds to avoid timeout.

Phylogenetic Distribution

You may view the phylogenetic distribution of best blast hits of protein-coding genes in selected scaffolds.

Distribution by BLAST percent identities

All of the genes in your genome have already undergone a BLASTx search. The phylogeny of the top hits for each gene in the scaffold is indicated. The scaffold shown below has 7 genes with a best hit of 90% identity or above to Gammaproteobacteria, 50 genes with a hit of 60%-90% to Gammaproteobacteria and 16 genes with top hits of 30%-60% to Gammaproteobacteria. In this case there are no genes with top hits to other phylogenetic groups.

img/mer **INTEGRATED MICROBIAL GENOMES**
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/ER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Marts Using

Home > My IMG/ER > Workspace > Scaffold Sets > f_class Loaded

Class Statistics in Selected Scaffolds (Gene Count)

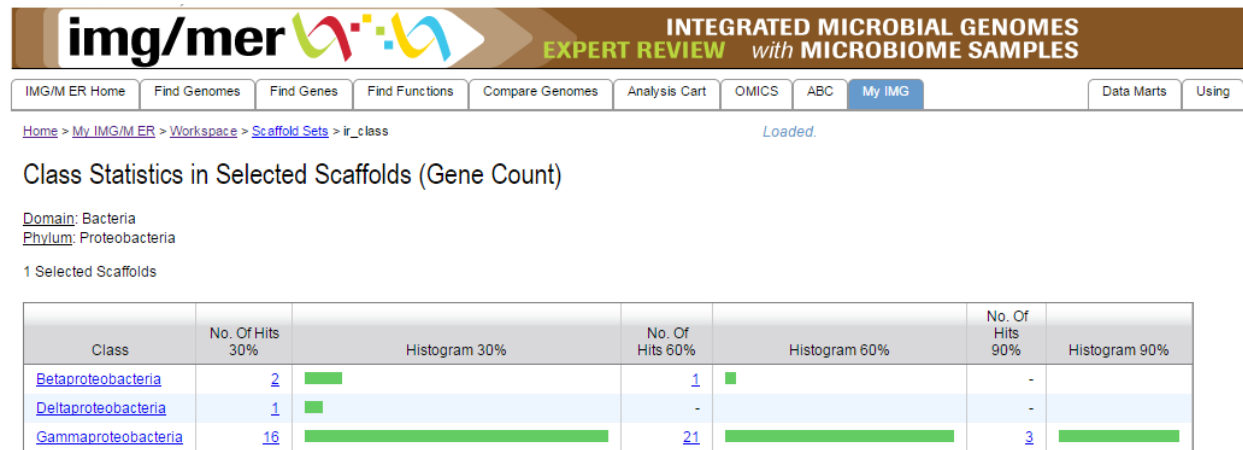
Domain: Bacteria
Phylum: Proteobacteria
1 Selected Scaffolds

Class	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%	Histogram 60%	No. Of Hits 90%	Histogram 90%
Gammaproteobacteria	16		50		7	

Click here to break the hits down by the next taxonomic level. (In this case to the orders.)

Click here to get a list of the genes and their top matches.

Due to gaps in the database as well as errors, you shouldn't place too much trust in a single gene's top hit. Instead, consider the consensus of all the genes on the scaffold before deciding whether it belongs to your genome or is a contaminant. For example, the scaffold below has three genes with best hits to Betaproteobacteria and one with a best hit to Deltaproteobacteria, but the majority of genes match Gammaproteobacteria and thus this scaffold can be confidently classified as coming from a Gammaproteobacterium.



There are no strict rules about what should be removed as a contaminant. You will have to decide based on your data and on your own tolerance for mistakenly leaving a contaminant in or mistakenly throwing out something that really belongs.

Ideally every scaffold would be checked individually. However, we recognize that that could take a lot of time and may not be necessary. Instead, you can focus on the subset of scaffolds that are suspicious. There are three main methods for identifying suspect scaffolds that need to be checked: scaffold GC%, Kmer analysis, and the phylogenetic distribution of the genes in all your unchecked scaffolds.

To identify scaffolds containing anomalous GC contents, go to the Scaffold Cart.

Select all the scaffolds and then go to the Histogram tab.

img/mer EXPERT REVIEW with MICROBIOME SAMPLES

Home > Analysis Cart 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis Phylogenetic Distribution

Add Genomes of Selected Scaffolds to Cart Add Genomes of Selected Scaffolds To Cart

Toggle Selected Select All Clear All Remove Selected

Filter column: Scaffold ID Filter text Apply

Export Page 1 of 1 < prev 1 next > last >> All

Column Selection Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05: Ga0063151_NODE_60_len_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	3	2079	0.30	1
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05: Ga0063151_NODE_22_len_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05: Ga0063151_NODE_43_len_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	8	5020	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05: Ga0063151_NODE_37_len_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05: Ga0063151_NODE_40_len_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	9	5946	0.36	1

img/mer EXPERT REVIEW with MICROBIOME SAMPLES

Home > Analysis Cart 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis Phylogenetic Distribution

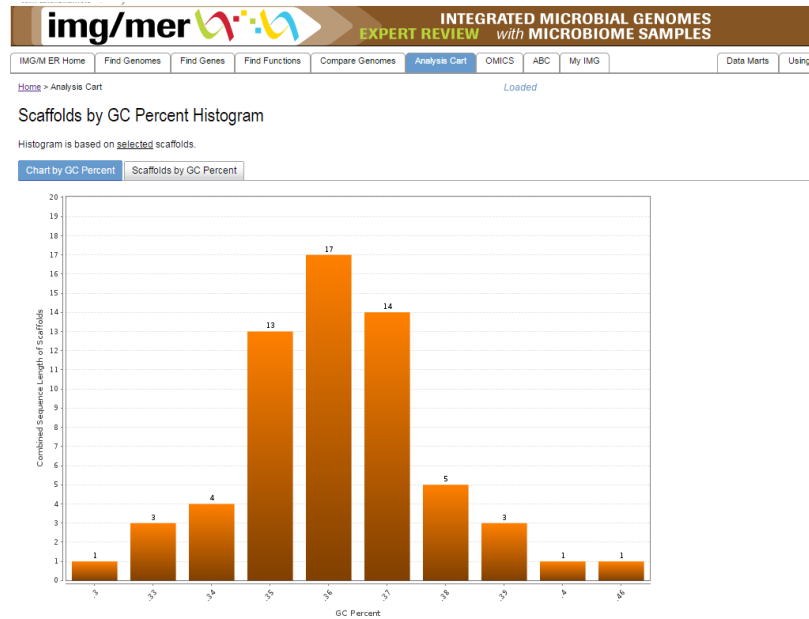
Histogram

You may compare selected scaffolds by:

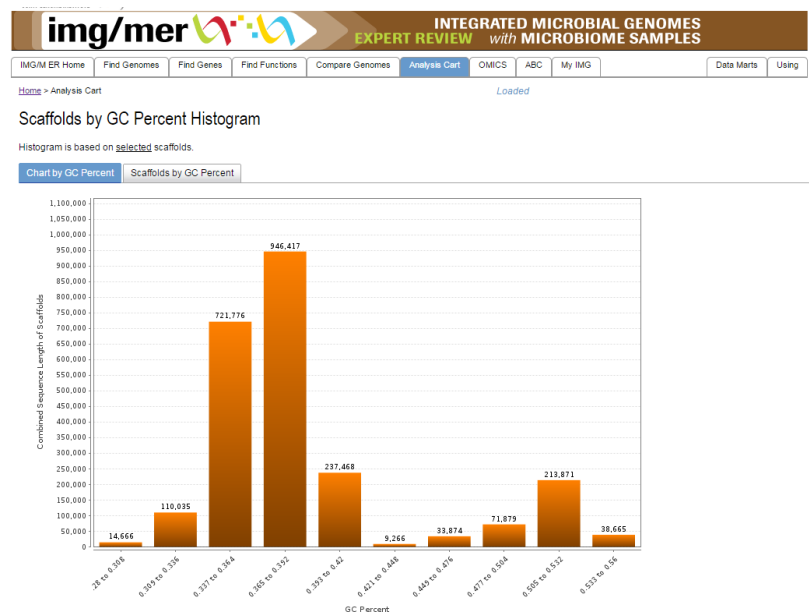
Gene Count
Gene Count
Sequence Length
GC Content
Read Depth

Show Histogram

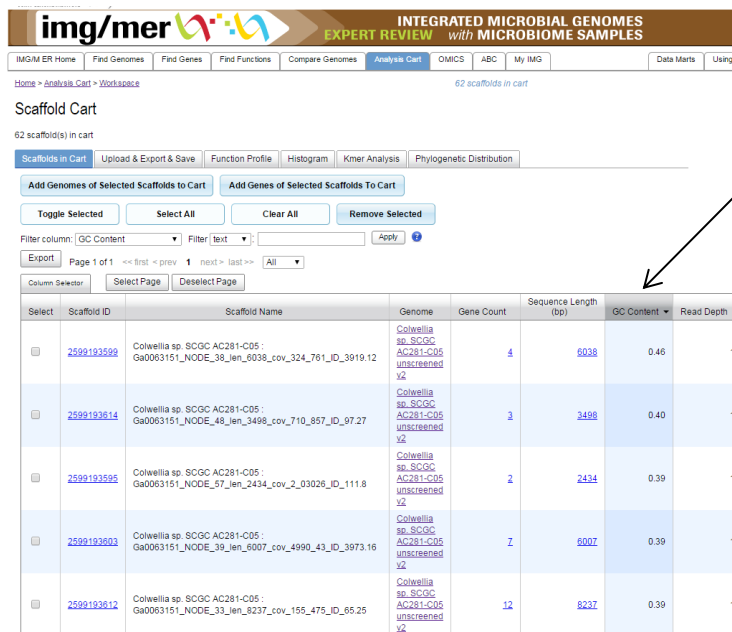
Select GC Content from the dropdown and then click the Show Histogram button.



For a clean genome, you should have a single peak and all scaffolds should be within ~10% to either side of the center. Note that the histogram always gives you 10 bars/bins no matter how wide the spread in GC content is and the bins are not necessarily contiguous. For the above genome the bar to the far right (46% GC) is significantly higher than the next one (40%) and should be investigated. An example of a highly contaminated genome is shown below.



Go back to the Scaffold Cart and sort by Scaffold GC% by clicking the top of the column.



img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/M ER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Maps Using

Home > Analysis Cart > Workspace 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis Phylogenetic Distribution

Add Genomes of Selected Scaffolds to Cart Add Genes of Selected Scaffolds To Cart

Toggle Selected Select All Clear All Remove Selected

Filter column: GC Content Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193599	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_38_len_6038_cov_324_761_ID_3919.12	Colwellia sp. SCGC AC281-C05 unscreened v4	4	6038	0.46	1
<input type="checkbox"/>	2599193614	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_48_len_3498_cov_710_857_ID_97.27	Colwellia sp. SCGC AC281-C05 unscreened v4	3	3498	0.40	1
<input type="checkbox"/>	2599193595	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_57_len_2434_cov_2_03026_ID_111.8	Colwellia sp. SCGC AC281-C05 unscreened v4	2	2434	0.39	1
<input type="checkbox"/>	2599193603	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_39_len_6007_cov_4990_43_ID_3973.16	Colwellia sp. SCGC AC281-C05 unscreened v4	2	6007	0.39	1
<input type="checkbox"/>	2599193612	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_33_len_8237_cov_155_475_ID_65.25	Colwellia sp. SCGC AC281-C05 unscreened v4	12	8237	0.39	1

Now you can select the individual scaffolds with suspect GC% and check them using the Phylogenetic Distribution of Genes as described earlier. Bin these scaffolds to the clean or contaminant workspace and remove them from your Scaffold Cart.

To identify suspect scaffolds by Kmer analysis, go back to the Scaffold Cart, select all the scaffolds and click on the Kmer Analysis button.

img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

Home > Analysis Cart > Workspace 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram **Kmer Analysis** Phylogenetic Distribution

Add Genomes of Selected Scaffolds to Cart Add Genes of Selected Scaffolds To Cart

Toggle Selected Select All Clear All Remove Selected

Filter column: Scaffold ID Filter text: Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector Select Page Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05 : Gao063151_NODE_60_1en_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	2	2079	0.30	1
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05 : Gao063151_NODE_22_1en_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05 : Gao063151_NODE_43_1en_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	6	5026	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05 : Gao063151_NODE_37_1en_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05 : Gao063151_NODE_40_1en_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	9	5946	0.36	1

img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

Home > Analysis Cart > Workspace 62 scaffolds in cart

Scaffold Cart

62 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram **Kmer Analysis** Phylogenetic Distribution

Scaffold Consistency Check

You may analyze selected scaffolds for purity using Kmer Frequency Analysis.

[Kmer Frequency Analysis](#)

img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

Home > Find Genomes loaded

Kmer Frequency Analysis

Change Kmer Settings

Lowering the 'Oligomer size' helps avoid memory issues

Parameter	Setting
Fragment window (1000 - 10000)	10000
Fragment step (100 - 1000)	500
Oligomer size (2 - 8)	4
Minimum variation (1 - 20)	10

Generate

Begin by sticking with the defaults. The larger scaffolds have more statistical power, which will produce a more defined cloud of points. Also, it is easier to get a feel for the data with the few large scaffolds than if you included all of the data. Later you will want to rerun this analysis with a smaller fragment window to include all your scaffolds in the screen.

Clicking the Generate button will produce a Kmer plot.

Kmer Frequency Analysis

Lowering the "Oligomer size" helps avoid running out of memory
Fragment window: 5000 bp, Fragment step: 500 bp, Oligomer size: 4, Minimum variation: 10

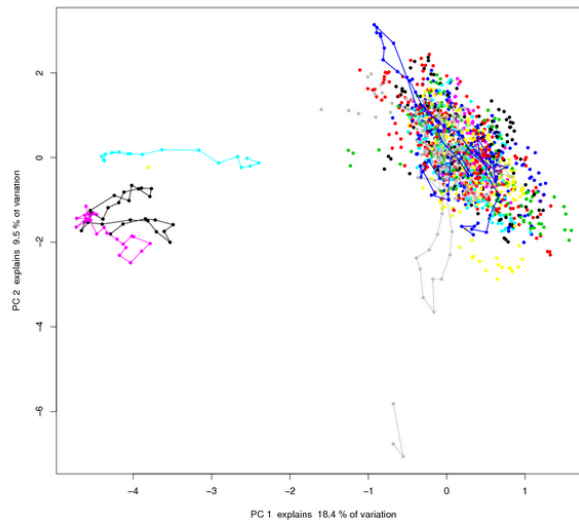
Change Settings

2D View

3D View

Mouse over a point to see the scaffold which it represents.
Click on a point to go to the Chromosome Viewer.
If item and/or .xml files were not created, the system could have run out of memory during computation - you may try to lower the "Oligomer size" and recompute.

Selected Scaffolds - PC1 vs PC2



There is both a 2D and a 3D view. Often the 3D view is more useful, though in this case the outliers are quite clear in the 2D view.

Click on the 3D View tab.

Kmer Frequency Analysis

Lowering the "Oligomer size" helps avoid running out of memory
Fragment window: 5000 bp, Fragment step: 500 bp, Oligomer size: 4, Minimum variation: 10

Change Settings

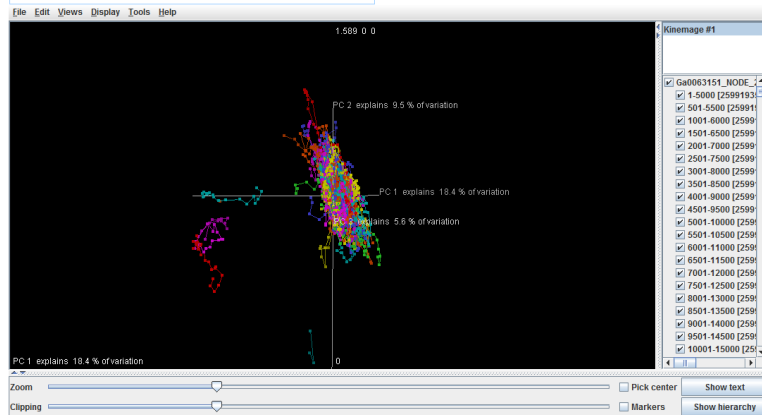
2D View

3D View

Selected Scaffolds - Plot of PC1, PC2, and PC3

The 3D view below is generated using the [3Dviz](#) applet.

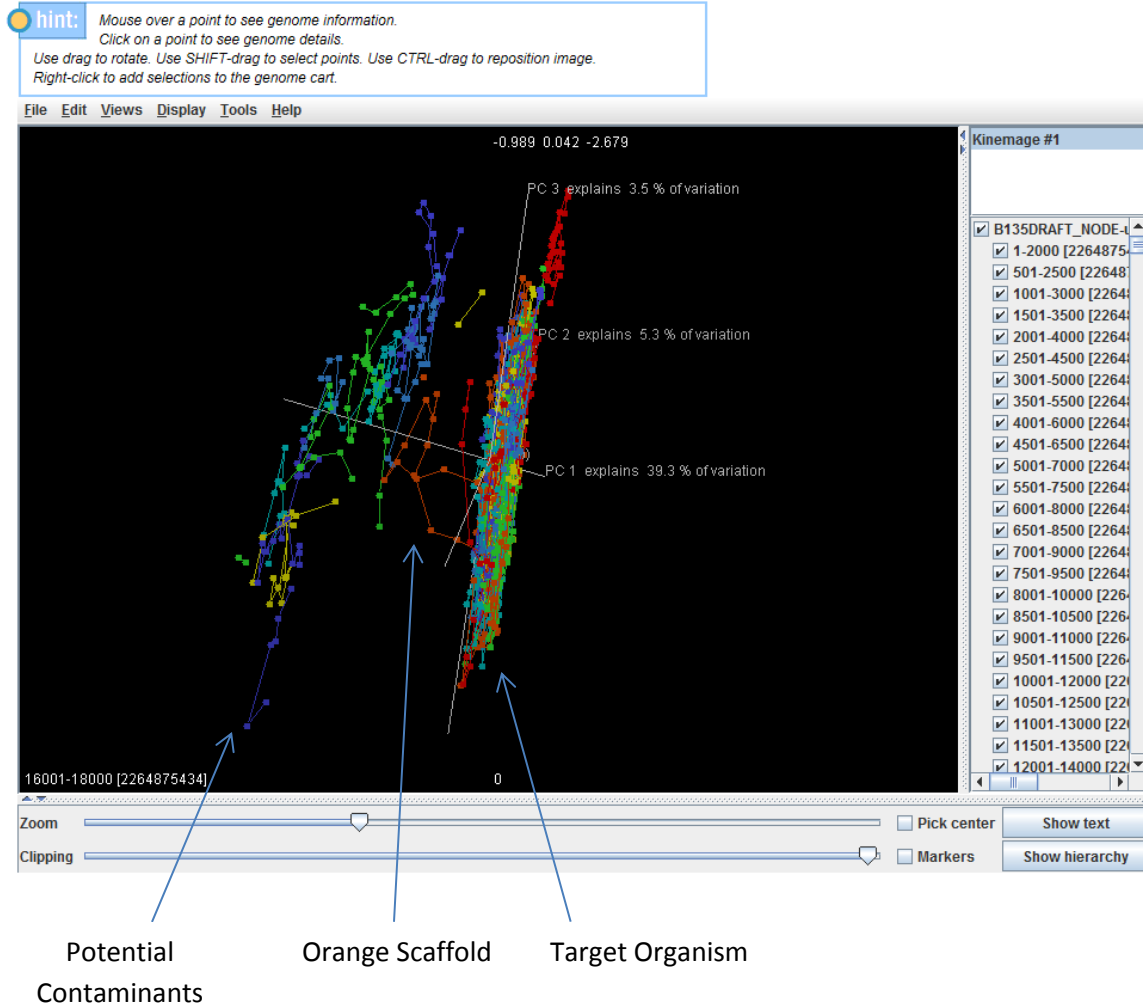
hint: Mouse over a point to see item information.
Click on a point to see item details. If the tooltip for a point of interest does not show the coordinates, try rotating the plot until the coordinates appear.
Use drag to rotate. Use SHIFT-drag to select points. Use CTRL-drag to reposition image.
Right-click to add selections to the cart.



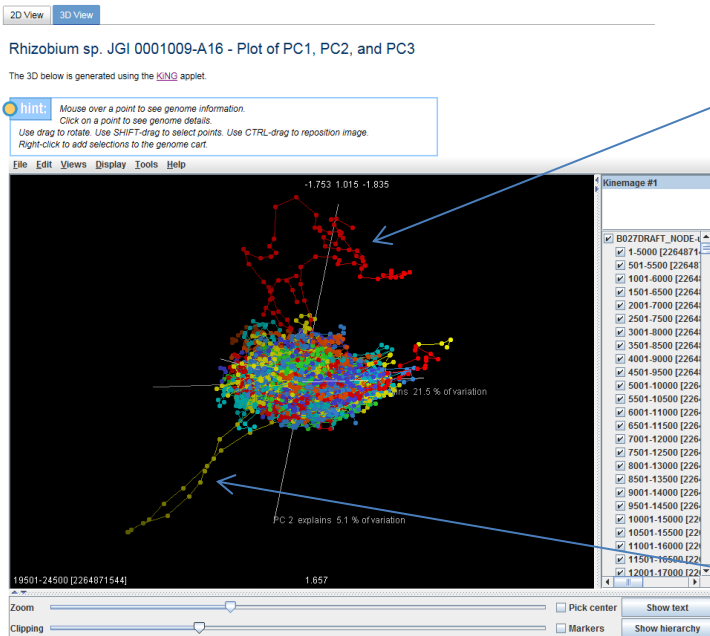
You can click and drag on the image to rotate it in three dimensions. First, look at the percent of variation explained by each principal component. If the percentages are all small (<~5%) then you have a very clean genome and the outliers are unlikely to be a problem.

Below is a fairly contaminated genome. Most points are in a large mass which is our target genome, but there is a distinct cloud of contaminant scaffolds to the left of the main cloud. By clicking on any of the points in the plot it will open a separate window of that scaffold.

The 3D below is generated using the [KING](#) applet.



Note the Orange scaffold. This one starts in the main cloud, extends into the contaminant zone, and then returns to the main cloud. Upon examining this scaffold, we find that the region that extends out from the main cloud contains rRNA genes that match the target organism. Ribosomal RNA genes often contain a different GC content from the rest of the genome and thus will plot outside the main cloud of your target genome. Scaffolds that extend from the main genome cloud can also contain other interesting features.



This red scaffold has points in the main cloud but extends well out.

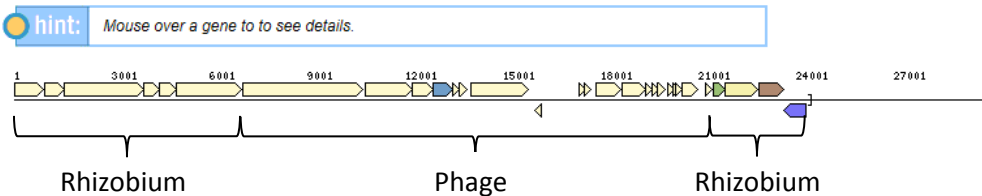
Clicking on the points in this scaffold opens a separate window with more detail on the scaffold shown below.

rRNA genes

Chromosome Viewer - Colored by COG

Switch coloring to:

Rhizobium sp. JGI 0001009-A16 : B027DRAFT_NODE-unique_54 len_24289.54 (bins: Rhizobium9A16 Cleaned(tetramer_GC)) (24289bp gc=0.58 depth=1.00) (coordinates 1-24289)



This scaffold is from a Rhizobium single cell and it contains 30 predicted genes. By doing a BLASTx search on each of the genes we found that the ends have high matches to proteins from various Rhizobium species. However, the genes in the middle, which caused the scaffold to stick out from the main cloud in the plot, have best matches to phage proteins. This cell appears to be infected with a lysogenic phage.

Suspicious scaffolds are those that lie outside the main cloud of points. You can identify these either by clicking on a point as above, or just by hovering the mouse over a point in the scaffold, the scaffold ID number appears in the bottom left corner of the plot.

Finally, you can identify suspicious scaffolds by looking at the Phylogenetic Distribution of Genes for all the unchecked scaffolds. Go back to your Scaffold Cart and Select All the scaffolds then click the Phylogenetic Distribution of Genes button.

84 scaffold(s) in cart

Buttons: Scaffold(s) in Cart, Upload & Export & Save, Function Profile, Histogram, Kmer Analysis, **Phylogenetic Distribution**

Buttons: Add Genomes of Selected Scaffolds to Cart, Add Genes of Selected Scaffolds To Cart

Buttons: Toggle Selected, Select All, Clear All, Remove Selected

Filter column: Scaffold ID | Filter text: | Apply

Export | Page 1 of 1 | << first < prev 1 next > last >> | All

Column Selector | Select Page | Deselect Page

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599194948	Cycloclasticus sp. SGC AC281-N15 : Ga0063159_NODE_81_len_2205_cov_1131_45_ID_2725.1	Cycloclasticus sp. SGC AC281-N15 unscreened v2	3	2205	0.41	1
<input type="checkbox"/>	2599194949	Cycloclasticus sp. SGC AC281-N15 : Ga0063159_NODE_82_len_2136_cov_2_73234_ID_173.2	Cycloclasticus sp. SGC AC281-N15 unscreened v2	3	2136	0.46	1

84 Selected Scaffolds

The Phylogenetic Distribution of Genes allows to assess the phylogenetic composition of a genome sample based on the distribution of best BLAST hits of protein-coding genes in the dataset. The phylogenetic distribution can be projected onto the families in a phylum (click on phylum name), and then further onto species in a family. For a reference genome within a species, the genome genes can be viewed using the Protein Recruitment Plot or the Reference Genome Context Viewer.

Buttons: Gene Count, Estimated Gene Copies

Distribution of Best Blast Hits (Gene Count)

Domains(D): * =Microbiome, B=Bacteria, A=Archaea, E=Eukarya, P=Plasmids, G=GFragment, V=Viruses.

hint: Hit genome count is in brackets (). Histogram is a count of best hits within the phylum / class at 30%, 60%, and 90% BLAST identities. Unassigned are the remainder of genes less than the percent identity cutoff, or that are not best hits at the cutoff, or have no hits.

D	Phylum	No. Of Genomes	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%	Histogram 60%	No. Of Hits 90%	Histogram 90%
B	Actinobacteria	2932 (2)	2 (2)					
B	Bacteroidetes	888 (6)	5 (4)		2 (2)			
B	Chlamydiae	162 (1)	1 (1)					
B	Chlorobi	14 (3)	2 (2)		1 (1)			
B	Chloroflexi	44 (1)	1 (1)					
B	Cyanobacteria	230 (6)	8 (6)					
B	Deinococcus-Thermus	58 (1)	1 (1)					
B	Firmicutes	8935 (7)	6 (6)		1 (1)		1 (1)	
B	Nitrospirae	14 (1)	4 (1)					
B	Planctomycetes	34 (2)	2 (2)					
B	Proteobacteria	10222 (197)	278 (149)		1079 (78)		278 (8)	
B	Spirochaetes	453 (1)	2 (1)					
B	unclassified	17 (1)	1 (1)					
E	Mollusca	3 (2)	1 (1)		1 (1)			
E	unclassified	23 (1)	1 (1)					
E	Annelida	2 (1)			1 (1)			
-	Unassigned	-	52		368		1453	

It looks like this genome may have some Firmicute and other miscellaneous contaminants.

Clicking here expands the Proteobacteria to show some potential contaminants in other Proteobacterial classes.

Class Statistics in Selected Scaffolds (Gene Count)

Domain: Bacteria
Phylum: Proteobacteria
84 Selected Scaffolds

Class	No. Of Hits 30%	Histogram 30%	No. Of Hits 60%	Histogram 60%	No. Of Hits 90%	Histogram 90%
Alphaproteobacteria	12		2		1	
Betaproteobacteria	18		8		-	
Deltaproteobacteria	8		5		-	
Epsilonproteobacteria	-		1		-	
Gammaproteobacteria	239		1053		277	
Zetaproteobacteria	1		9		-	
unclassified	2		1		-	

To identify which scaffolds contain these suspect genes, click on one of the numbers.

img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/MER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Maps Using

Home > My IMG/MER > [Workflow](#) > [Scaffold Sets](#) > taxonomy/metagene 8 gene(s) retrieved

Best Hits at 60% Identity

Domain: Bacteria
Phylum: Proteobacteria
Class: Betaproteobacteria

84 Selected Scaffolds

Filter column: Gene ID Filter: text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Select	Gene ID	Name	Percent	Homolog Gene	Homolog Genome	Homolog Class	Homolog Order	Homolog Family	Homolog Genus	Homolog Species
<input type="checkbox"/>	2599907849	Acetyl-CoA acetyltransferase [Cycloclasticus sp. SCGC AC281-N15 unscreened v2]	76.70	2514530485	Acidovorax sp. NO-1	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax	Acidovorax sp. NO-1
<input type="checkbox"/>	2599908018	Uncharacterized conserved protein [Cycloclasticus sp. SCGC AC281-N15 unscreened v2]	76.30	2592587799	Betaproteobacteria bacterium MGLA814	Betaproteobacteria	unclassified	unclassified	unclassified	Betaproteobacteria bacterium MGLA814
<input type="checkbox"/>	2599908146	DegT/OmrA/EryC1/SrS aminotransferase family [Cycloclasticus sp. SCGC AC281-N15 unscreened v2]	69.20	2530790496	Comamonas testosteroni ATCC 11996	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonas	Comamonas testosteroni

This is a list of the 8 genes that had a best hit that was 60-90% to a Betaproteobacteria.

This list does not include the scaffolds that each gene is found on. To get that information scroll down.

<input type="checkbox"/>	2599909594	hypothetical protein [Cycloclasticus sp. SCGC AC281-N15 unscreened v2]	61.30	2515446492	Thiobacillus thioautotrophicus DSM 595	Betaproteobacteria	Hydrogenophiales	Hydrogenophiales	Thiobacillus	Thiobacillus thioautotrophicus
<input type="checkbox"/>	2599909591	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) [Cycloclasticus sp. SCGC AC281-N15 unscreened v2]	66.80	639839177	Acidovorax sp. JS42	Betaproteobacteria	Burkholderiales	Comamonadaceae	Acidovorax	Acidovorax sp. JS42

Export Page 1 of 1 << first < prev 1 next > last >> All

Add Selected to Gene Cart Select All Clear All

Expand Gene Table Display

Limit gene selection and display options to avoid timeout.

Display Options for Selected Genes:

- ☐ COG Alignment
- ☐ Pfam Alignment
- ☒ Display by including the following information
 - ☒ Gene Detailed Information
 - ☒ Scaffold Information (for assembled only)
 - ☐ COG Functions
 - ☐ Pfam Functions
 - ☐ TrnGene Functions
 - ☐ Enzyme Functions
 - ☐ K0 Functions

Go

Select all the genes, check the box indicating that you want their Scaffold Information and then click the Go button.

img/mer INTEGRATED MICROBIAL GENOMES
EXPERT REVIEW with MICROBIOME SAMPLES

IMG/MER Home Find Genomes Find Genes Find Functions Compare Genomes Analysis Cart OMICS ABC My IMG Data Maps Using

Home > Find Genes 8 gene(s) loaded

Expanded Gene List

Filter column: Gene ID Filter: text Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Select	Gene ID	Gene Name	Taxon ID	Assembled?	Locus Type	Start Coord	End Coord	Gene Length	Strand	Scaffold	Scaffold Length	Scaffold GC	Scaffold Depth	# of Genes on Scaffold
<input type="checkbox"/>	2599907849	Acetyl-CoA acetyltransferase	2599185293	assembled	CDS	1061	2134	1074	-	2599194949	2136	0.46	1	3
<input type="checkbox"/>	2599908018	Uncharacterized conserved protein	2599185293	assembled	CDS	3819	5588	1770	+	2599194962	30794	0.43	1	24
<input type="checkbox"/>	2599908146	DegT/OmrA/EryC1/SrS aminotransferase family	2599185293	assembled	CDS	16059	16631	573	+	2599194969	17376	0.46	1	13
<input type="checkbox"/>	2599908439	AAA domain	2599185293	assembled	CDS	16246	16746	501	-	2599194985	56491	0.42	1	56
<input type="checkbox"/>	2599909127	addiction module antibiotic protein, HigA family	2599185293	assembled	CDS	36640	36921	282	-	2599195017	131426	0.42	1	128
<input type="checkbox"/>	2599909593	Predicted pyridoxal phosphate-dependent enzyme apparently involved in regulation of cell wall biogenesis	2599185293	assembled	CDS	521	1621	1101	+	2599195028	12019	0.43	1	11
<input type="checkbox"/>	2599909594	hypothetical protein	2599185293	assembled	CDS	1615	2562	948	+	2599195028	12019	0.43	1	11
<input type="checkbox"/>	2599909591	Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases)	2599185293	assembled	CDS	16167	16922	756	+	2599195030	17097	0.42	1	16

Export Page 1 of 1 << first < prev 1 next > last >> All

Add Selected to Gene Cart Select All Clear All

These are the IDs for the scaffolds that you will want to check.

Once you have finished cleaning up your genome you will want to upload it to IMG. The IMG system does not have a direct way to upload the data so first you must download a fasta file of your cleaned genome.

First, make sure your Scaffold Cart is empty by selecting any remaining scaffolds and removing them. Next, go to your workspaces under the My IMG tab.

Select the workspace that contains all your clean scaffolds and add them to the Scaffold Cart.

Select All the scaffolds and go to the Upload & Export & Save tab.

Select	Scaffold ID	Scaffold Name	Genome	Gene Count	Sequence Length (bp)	GC Content	Read Depth
<input type="checkbox"/>	2599193588	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_60_len_2079_cov_8155_82_ID_153.1	Colwellia sp. SCGC AC281-C05 unscreened v2	3	2079	0.30	1
<input type="checkbox"/>	2599193589	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_22_len_18811_cov_4213_38_ID_767.2	Colwellia sp. SCGC AC281-C05 unscreened v2	19	18811	0.37	1
<input type="checkbox"/>	2599193590	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_43_len_5027_cov_4_02635_ID_79.3	Colwellia sp. SCGC AC281-C05 unscreened v2	6	5026	0.38	1
<input type="checkbox"/>	2599193591	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_37_len_6434_cov_3418_15_ID_151.4	Colwellia sp. SCGC AC281-C05 unscreened v2	13	6434	0.35	1
<input type="checkbox"/>	2599193592	Colwellia sp. SCGC AC281-C05 : Ga0063151_NODE_40_len_5946_cov_417_481_ID_81.5	Colwellia sp. SCGC AC281-C05 unscreened v2	2	5946	0.36	1

Home > Analysis Cart

58 scaffolds in cart

Scaffold Cart

58 scaffold(s) in cart

Scaffolds in Cart Upload & Export & Save Function Profile Histogram Kmer Analysis Phylogenetic Distribution

Upload Scaffold Cart

You may upload a scaffold cart from a tab-delimited file.
The file should have the column headers 'Scaffold ID'.
(This file can be created by selecting [Scaffold Cart in Excel](#) button below.)

File to upload:

Choose File | No file chosen

Upload from File

Upload from Workspace

Export Scaffold Data


You may export data for scaffolds selected in the cart.

Fasta Nucleic Acid File

Genbank File

Scaffold Cart in Excel

Save Scaffolds to My Workspace

 **hint:** Even though you can save large amount of data into workspace, many profile functions will timeout for extremely large workspace datasets

Save selected scaffolds to [My Workspace](#).

(Special characters in file name will be removed and spaces converted to _)

Save to File name:

Append to the following scaffold set:

Replacing the following scaffold set:

contaminants

Save Selected to Workspace

Export all the data as a
Fasta Nucleic Acid File.

Copy all the data and save it on
your computer.

```
>2599193588 Ga0063151_N00E_60_1en_2079_cov_8155_82_ID_153.1
TCAACG6GACAAATAACAGTGGGCTGCCGTCTGCGCTCACTATTTTAGC
CAACTATCTATTTGCCGCTTAACGCGCGCTTATGCATTAGTATCTGAATTT
AGCTTTTATGTGATGAATTTATCTTAAAGCTTCATGAATATATCTTG
TTCTTAATGTTTACTTTAACAATATTAATTTGATCAACAATGTTGATT
GGCTTATTAATTTTAAATTAATCTATCTTTAAGGAATAATTTTCATGAA
AAATAAATATATAAATGGAAGTCAACTGTAGCTACAGTATTTTATAGCAA
CATCTCTTATTTCTAGCTCTATATTGCTAGTGATTAAGGGGGAGATA
ATAGACGTTGAGCCCAACAATTATTAAGGATACACGATTAATGGAATTC
TTAGATTTATCAGATTTAAGTCCATGAATGAAGATCAAAAGTTCAAG
TTTCTACAAAATCCGCGGACAGGTTGATTATTTGAAATATAGCT
GTTGGATCATCAATGTTGGATGGGAATGGTTTCTCTCAATCAGAGTTT
AACATCATATAACCATGGAGGAAACAATTAAAGATGGCTGTTTACAAA
TAGGGTTAGGTAATCTCAATATGAACCATGGGAGGTGTCAATACATCT
AATTATGCTAGTGAATTTATGTTGGCTCTAATTTACATATATGTAATTC
TGGTGAATAGTTACAGGCTTTTACGCTACTATCTTTTGTAGGTCAC
AAAGCGGTATTTTCAAACTCAACGAATCTGTTGCTAGCCCTTTGGC
TATTGGTCTGATCTATATCAATCAATAGATATAAATTTTAAAAAG
CCGTATTTATTAATAGCGGTTTATTTAAATGATCTTTATTTGGCTTAA
TAATGCTTATTTATATTGTTGCTGTTAAAGTTTGGAGGATTAATGAT
ATTTCAAAATAGGGCAAAATCTGCTGTTTATTTATGAAGGCGCA
TAATTTACCGTTACAAGTTAATGAAGTGAAGGACACAGATGATATAT
CTCAACCTCAAAAAAATGATCTGAAAAATATTCAATGATGAAATA
AATGGATTAATCCGTTCTGGTGTGATGGTGGTTTATTGGATTTTATACC
AATAAGGCTCAATATTAGATATATATGGAAGTTAATGAAGGAGCAGGTT
ATCTCTCAATGAAGAAATAGATCTTAAAGTCAAGTTAAAGAGGATTA
GAGTTTCAAAAAGCATAGGAAGCACTGATTTTTGAAGAAATACAT
AAAAAATACACTGAGCTGTGAGGACGTTATTTCTCAAAAAATAGATT
TAACAGCTTAATACTATAATAGTAGATGATGAAGGAAATGAACATAG
GAAAAATAGAGTTTGCCGATTTTGTCAAAAGGCGCAAAAAAGCTCCA
ACTCAATTTTCTGTATTTAAAGCTTAAGGCTTTTCCGCACTCAATAG
TTAGTTTATGTTAATAGGTTGGTGGAGGAAATGAAGGAAATCATAGT
GTTAATATCTTTTATGTTGACAGTTGTGCTAGTTAGAAAGTGCA
GTCAATATTTAAAGAACAGGACGAGAAATAGGCTTAATCTTATAT
CAAGTGTAGATGAAGATTTAATGATCATTTTATAAATATCAAAAA
CGAGAGAGGCTGTTTAAAGTGAATGAATGAAGATGATTTTCTT
TTGAAGGCTTGAGGCTCTAGATATATCTTAAAGGATCTATGATAT
GGAGGAAATGGGGCAATATCAAAAAATGATCAATATTCATACAA
TAGTGAAGTGAATGTTGCTAAAAATATCAAAATGGGAGTAAAA
GAGGAAATGTTGTTAAATATACAGGCTAGATTAACAAAGTAATTAA
GGATTATTGAAGCAACATTTTCAAGCAGAAAAAATCAAGGAGCATGTA
CGGTATTGATATAGTGAATGAATGAATGAATGAATGAATGAATGA
CTTTTTTAGTAGAGGACACCAAGCAAGCA
```

```
>2599193589 Ga0063151_N00E_22_1en_18811_cov_4213_38_ID_767.2
AGAATTTCTTACCCATGTGCTATAGAAGACAGTCAATACATATAGAGC
ATTATCAAGAGTCTTAAGAGGTTAAATTAATCAACACATATGAT
GCAGGTGACAGTGTGTTGGTGTCTATCTGACACTAGTAGCAAGTAATA
CATTGCTATAGCAAGTAATATCAATCAACAAATTTAATTTACCCGA
GTGTTGATTACACTTTTCACTCCCTCTATCGATGAAATGGGACAGGG
TTCTTTTGAAGAAAGACAAATTTGGCTGATTTTAACTATTCTTTCA
AGCAATGAAGTTGAGGACATGCTCACTTTGTTATGCTCAATGGAG
CTAACATGCCAGAAACAATTTTACGCGCGTTGCGATCCGCTTAGA
GATGAAGGCTGCTATGCGGATGATTAACAATGGGTTTAAAGT
AGAACAGCATACCTTTGAGGGTATGATCATGCTTATATGATTTAGATA
GCTTAGTACCGCAAGATGTGAAGAACCTATCAGATGATTTGGTGAATTT
ATTGCAACATTAATCTTTTAACTTTTAACTTTGTTGATCTGAAGGATGAAGC
AAGCTATTAAATCTGACAGCGGTTGCTGTTTCAAAATCAACAGCGTA
CTTTTTGACTAACTACCGGTGATGGAATGATTAAACCATTTGTTA
```

In order to upload this clean dataset to IMG you first need to get an Analysis Project ID from the GOLD website at www.genomesonline.org.

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

GOLD
Genomes Online Database

Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News

Studies 20571
Biosamples 60669
Sequencing Projects 60651
Analysis Projects 47019


[Download Excel Data file](#)

Welcome to the Genomes OnLine Database
GOLD Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

GOLD Release v.5

Studies	Biosamples	Projects	Organisms
• Metagenomic 553 • Non-Metagenomic 20025	• Classification • Ecosystems • Host-associated 11910 • Engineered 1669 • Environmental 6891	• Complete Projects 6651 • Permanent Drafts 23543 • Incomplete Projects 28411 • Targeted Projects 1253	• Organisms 68151 • Archaea 1037 • Bacteria 44576 • Eukarya 8181


1. Register



Register your project information and Metadata in the Genomes Online Database

[Register](#)


2. Annotate



Annotate your microbial genome or metagenome with IMG/ER or IMG/MER

[Annotate](#)

3. Publish



Standards in Genomic Sciences

Publish your genome or metagenome in open access standards-supportive journal.

[Publish](#)

Please cite:
Reddy TBK, Thomas A, Stamatis D, Bertsch J, Isbrandt M, Jansson J, Mallajosyula J, Pagani I, Lobos E and Kyriides N. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucl. Acids Res.* (2014) doi: 10.1093/nar/gku950
[Full text](#)

©2014 The Regents of the University of California
[Disclaimer](#) | [Credits](#)

U.S. DEPARTMENT OF **ENERGY** | Office of Science

Click Register and then Create a new Analysis Project for submission to IMG.

JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

GOLD
Genomes Online Database

Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News My Profile

Welcome, Scott Clingenpeel [Log out](#)

Studies 20571
Biosamples 60669
Sequencing Projects 60651
Analysis Projects 47019

Create Projects, Studies and Biosamples

For information on the new structure of GOLD please [review the GOLD Project Entry Help Document](#)

- ▶ Create a new Sequencing Project in GOLD
- ▶ Create a new Analysis Project for submission to IMG
- ▶ Review your Studies, Biosamples and Sequencing and Analysis Projects

My Data:

Studies	0
Biosamples	0
Sequencing Projects	0
Analysis Projects	0

©2014 The Regents of the University of California
[Disclaimer](#) | [Credits](#)

U.S. DEPARTMENT OF **ENERGY** | Office of Science

Once you have your Analysis Project ID, go to IMG to submit your dataset.

img/mer EXPERT REVIEW with MICROBIOME SAMPLES

The Integrated Microbial Genomes (IMG) system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The IMG data warehouse integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available isolate and single cell genomes and a rich set of publicly available metagenome samples.

IMG/M ER (Nucleic Acids Research, Volume 42, Issue D1) provides users with tools (IMG/M UI Map) for analyzing their private (password protected access) metagenome samples in the context of all public (free access) genome and metagenome samples in IMG.

IMG/M ER contains 245 public studies, 3413 public metagenome datasets (3199 unique samples) distributed as follows:

Engineered	205	Environmental	2052	Host-associated	1156
Bioreactor	16	Air	31	Annelida	34
Bioremediation	21	Aquatic	1209	Arthropoda	77
Biotransformation	26	Terrestrial	812	Birds	5
Food production	3			Cnidaria	2
Lab enrichment	18			Human	861
Solid waste	23			Mammals	27
Wastewater	98			Microbial	3
				Mollusca	9
				Plants	122
				Porifera	8
				Tunicates	8

2015-03-19-08:00:01

Select Submit Data Set from the Companion Systems tab from anywhere in IMG/MER, or click the Data Submission Site button on the home page.

Select New Submission.

img/er & img/mer EXPERT REVIEW DATA SUBMISSION MICROBIAL GENOMES & METAGENOMES

Submission Home Submitted Datasets New Submission Filter Statistics FAQ

IMG Databases
You can check your loaded genomes in IMG by clicking the following links:

- [IMG ER](#)
- [IMG/MER](#)

Report all problems to: [Amy Chen](#)

[Logout](#)
[Accessibility/Section 508](#)

The Microbial Genome & Metagenome Expert Review Data Submission site allows scientists to submit genome datasets to IMG ER or metagenome datasets to IMG/MER in order to analyze and curate them in the context of a large set of reference public genomes and metagenomes.

What is Provided

IMG provides free support for genome & metagenome data annotation & integration and open access comparative analysis of integrated genome and metagenomes.

The data release and distribution policies and metadata requirements (listed below) will be strictly enforced.

User Support

We are committed to support scientists worldwide with their genome and metagenome data analysis needs. Decreased funding for IMG forces us to reduce the support provided to non JGI users, that is users who do not use JGI for sequencing. We continue to accept non JGI genome & metagenome datasets for annotation & integration, but our ability to respond to user requests and questions will be substantially reduced.

Data Release

Genome and metagenome datasets submitted for annotation and/or integration in IMG will be kept "private" for up to two years from the date they become available for analysis; then they will become public: isolate genome datasets will be kept private for 18 months, while single cell and metagenome datasets will be kept private for 24 months.

A genome or metagenome dataset submitted to IMG can be replaced by newer versions of the same genome/metagenome dataset, but cannot be removed in order to avoid making them public.

Follow the instructions to upload your cleaned genome.

Congratulations! Once this is annotated and loaded into IMG you will have a single cell genome to analyze that is free of contamination sequences.